# FoundationDB

Evan Tschannen

# FoundationDB

Evan Tschannen

- Worked on FoundationDB for 8 years
- Touched every core component of the database
- Recently added multi-region asynchronous replication

# FoundationDB

- https://www.foundationdb.org
- https://github.com/apple/foundationdb
- Open Source, Apache License (v2)

# Why are there so many databases?

348 systems in ranking, November 2018

| Rank Nov 2018 | Rank Oct 2018 | Rank Nov 2017 | DBMS | Database Model | Score Nov 2018 | Score Oct 2018 | Score Nov 2017 |
|---|---|---|---|---|---|---|---|
| 1. | 1. | 1. | Oracle | Relational DBMS | 1301.11 | -18.16 | -58.94 |
| 2. | 2. | 2. | MySQL | Relational DBMS | 1159.89 | -18.22 | -162.14 |
| 3. | 3. | 3. | Microsoft SQL Server | Relational DBMS | 1051.55 | -6.78 | -163.53 |
| 4. | 4. | 4. | PostgreSQL | Relational DBMS | 440.24 | +20.85 | +60.33 |
| 5. | 5. | 5. | MongoDB | Document store | 369.48 | +6.30 | +39.01 |
| 6. | 6. | 6. | DB2 | Relational DBMS | 179.87 | +0.19 | -14.19 |
| 7. | 7. | ↑ 9. | Redis | Key-value store | 144.17 | -1.12 | +22.99 |
| 8. | 8. | ↑ 10. | Elasticsearch | Search engine | 143.46 | +1.13 | +24.05 |
| 9. | 9. | ↓ 7. | Microsoft Access | Relational DBMS | 138.44 | +1.64 | +5.12 |
| 10. | ↑ 11. | ↑ 11. | SQLite | Relational DBMS | 122.71 | +5.96 | +9.95 |
| 11. | ↓ 10. | ↓ 8. | Cassandra | Wide column store | 121.74 | -1.64 | -2.47 |
| 12. | ↑ 13. | ↑ 15. | Splunk | Search engine | 80.37 | +3.48 | +15.50 |
| 13. | ↓ 12. | ↓ 12. | Teradata | Relational DBMS | 79.31 | +0.67 | +1.07 |
| 14. | 14. | ↑ 18. | MariaDB | Relational DBMS | 73.25 | +0.12 | +17.96 |
| 15. | ↑ 16. | ↑ 19. | Hive | Relational DBMS | 64.57 | +3.47 | +11.32 |
| 16. | ↓ 15. | ↓ 13. | Solr | Search engine | 60.87 | -0.44 | -8.28 |
| 17. | 17. | ↓ 16. | HBase | Wide column store | 60.41 | -0.26 | -3.15 |
| 18. | 18. | ↓ 14. | SAP Adaptive Server | Relational DBMS | 56.57 | -2.00 | -10.47 |
| 19. | ↑ 21. | ↑ 20. | SAP HANA | Relational DBMS | 55.88 | +1.50 | +6.70 |
| 20. | ↓ 19. | ↓ 17. | FileMaker | Relational DBMS | 55.75 | -0.29 | -3.09 |
| 21. | ↓ 20. | ↑ 22. | Amazon DynamoDB | Multi-model | 53.81 | -0.65 | +16.69 |
| 22. | 22. | ↓ 21. | Neo4j | Graph DBMS | 43.12 | +0.47 | +4.67 |
| 23. | 23. | 23. | Couchbase | Document store | 34.85 | -1.06 | +2.54 |
| 24. | 24. | 24. | Memcached | Key-value store | 29.75 | -0.80 | +1.77 |
| 25. | ↑ 26. | 25. | Informix | Relational DBMS | 26.45 | +0.21 | -1.26 |
| 26. | ↓ 25. | 26. | Microsoft Azure SQL Database | Relational DBMS | 26.19 | -0.08 | +4.08 |
| 27. | ↑ 28. | ↑ 32. | Microsoft Azure Cosmos DB | Multi-model | 22.03 | +1.78 | +9.00 |
| 28. | ↓ 27. | ↓ 27. | Vertica | Relational DBMS | 21.01 | -0.35 | -0.63 |
| 29. | 29. | ↑ 30. | Firebird | Relational DBMS | 20.52 | +0.53 | +3.24 |
| 30. | 30. | ↓ 28. | CouchDB | Document store | 18.73 | -0.66 | -1.78 |

# Why are there so many databases?

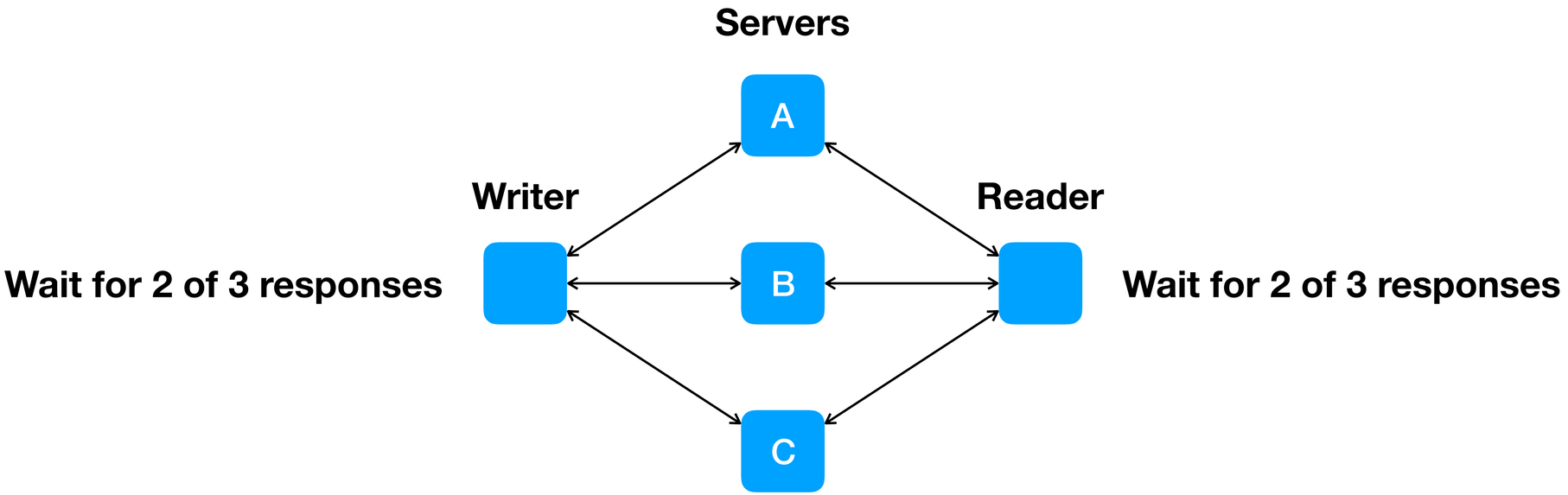
- Tension between performance, scalability, and consistency
- A wide variety of data models

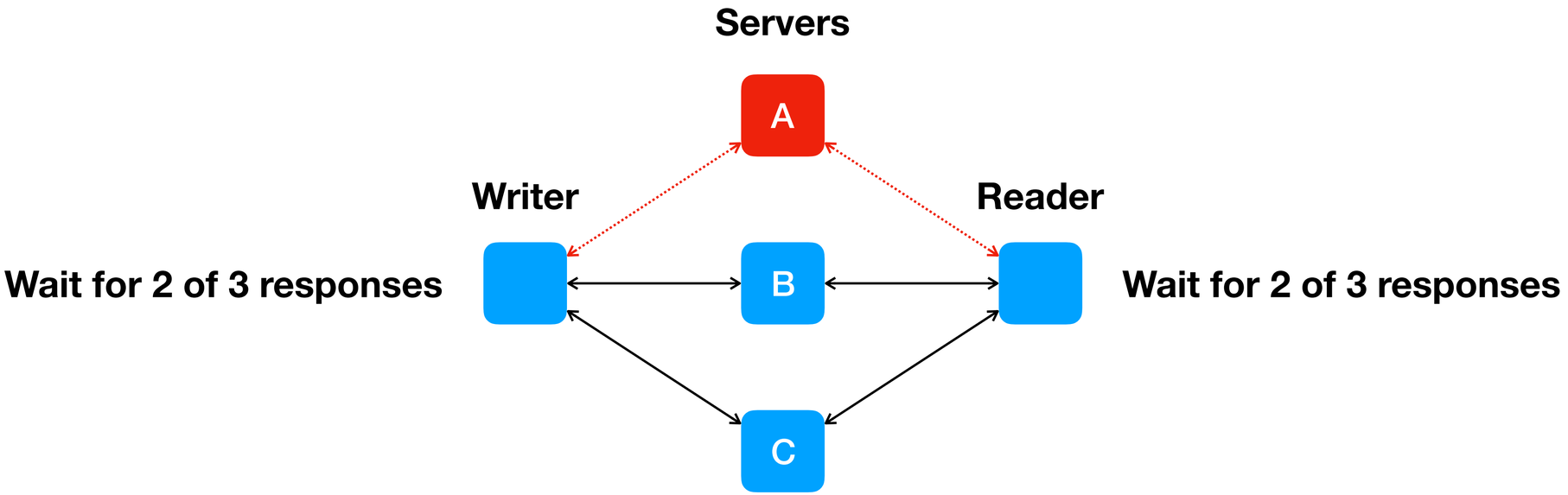
# FoundationDB Performance

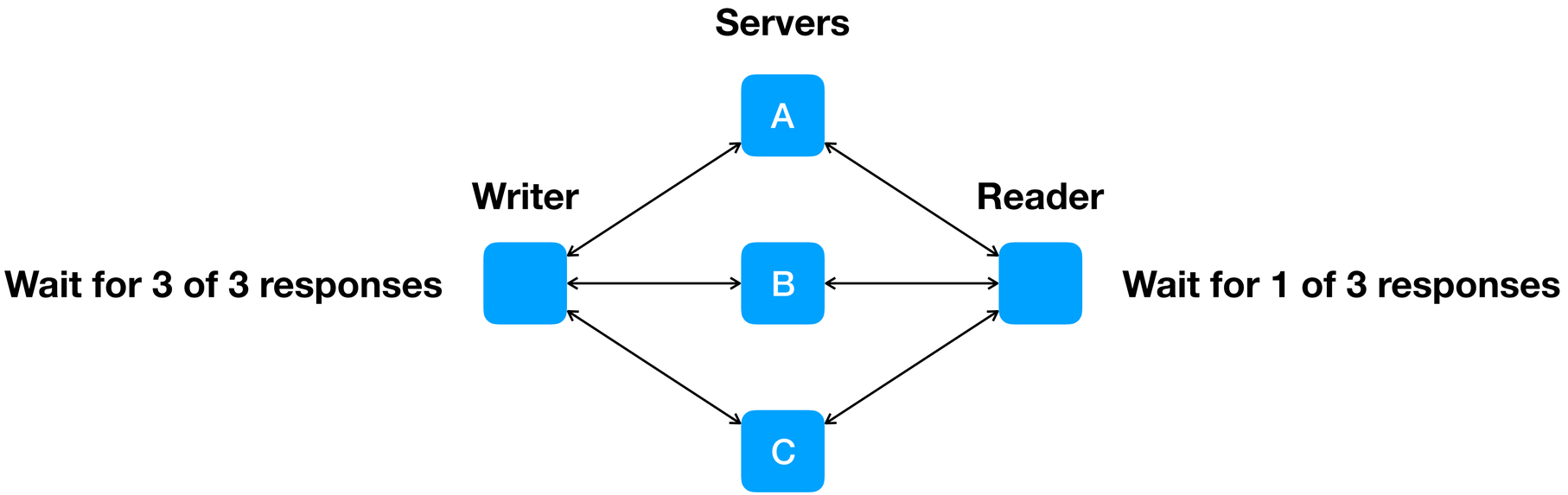
- Organizes many instances of a single process database into a single database
- Read and write throughput 90% of the aggregate individual read and write throughput
- Single hop read latencies
- Four hop write latencies

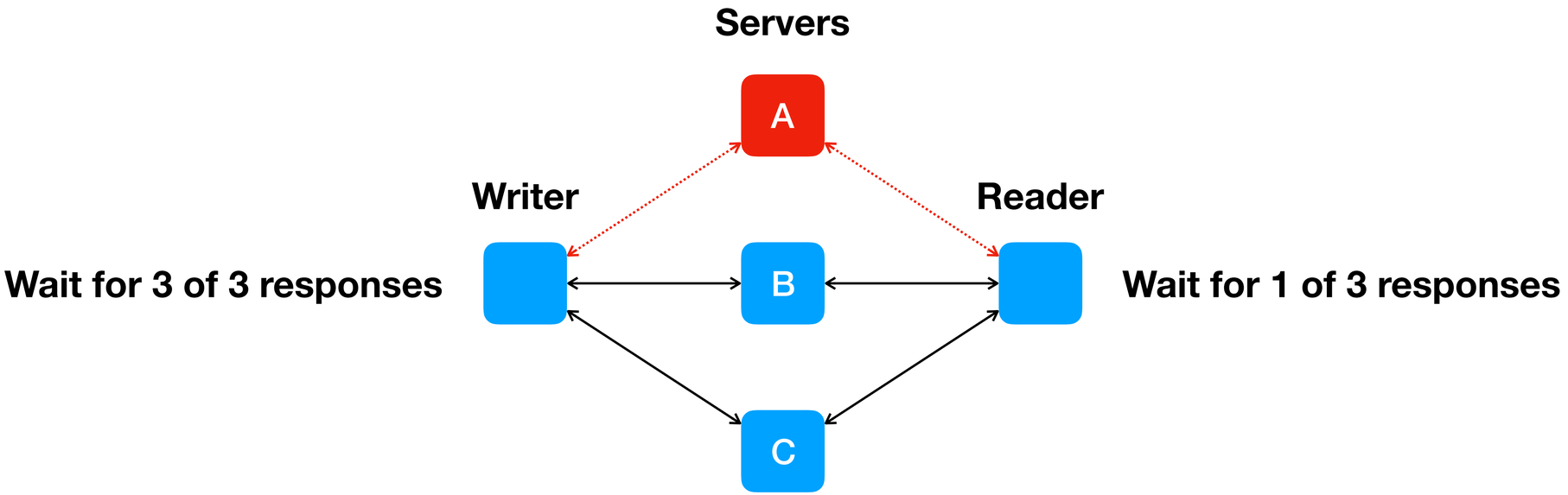
# FoundationDB is a…

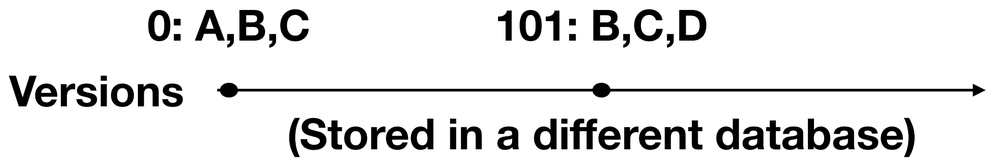
- Highly performant,
- Scalable,
- Ordered key value store,
- That supports ACID transactions.

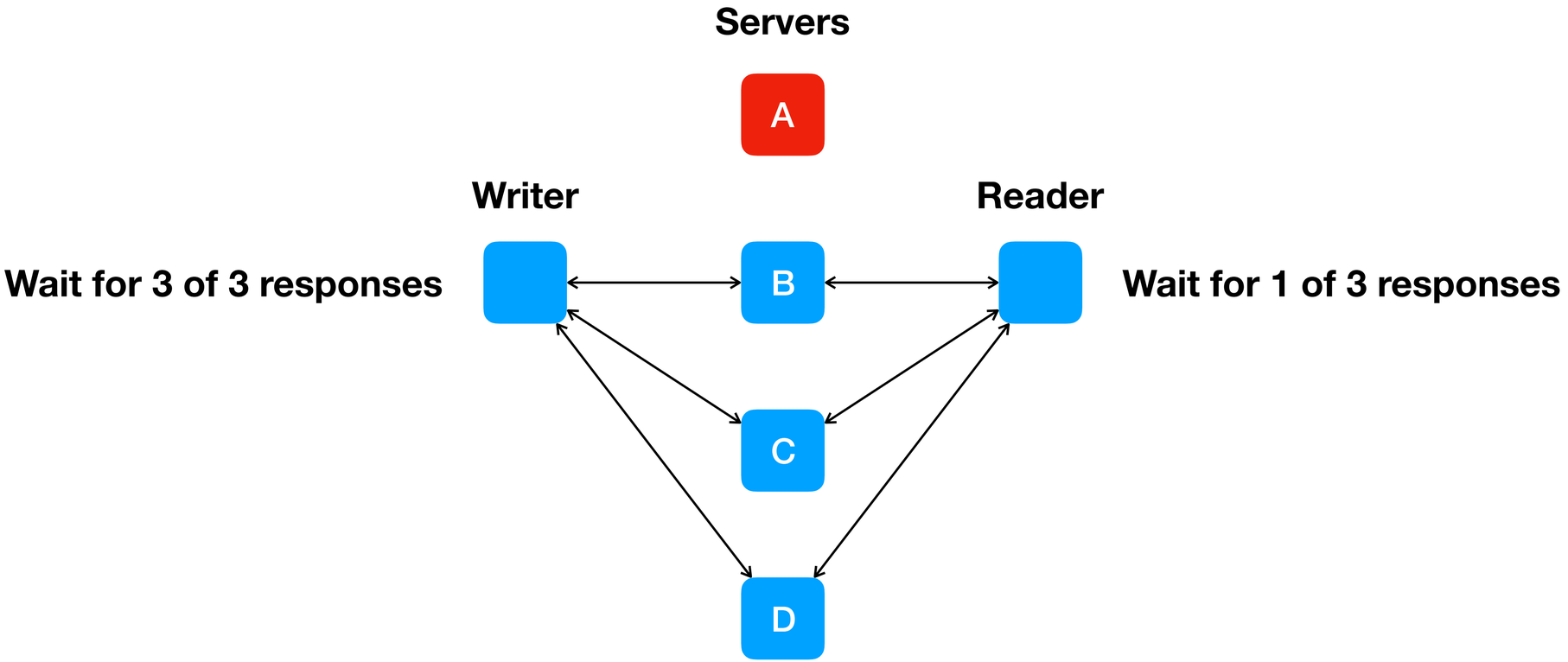

Servers
A
Writer
Reader
Wait for 2 of 3 responses
B
Wait for 2 of 3 responses
C

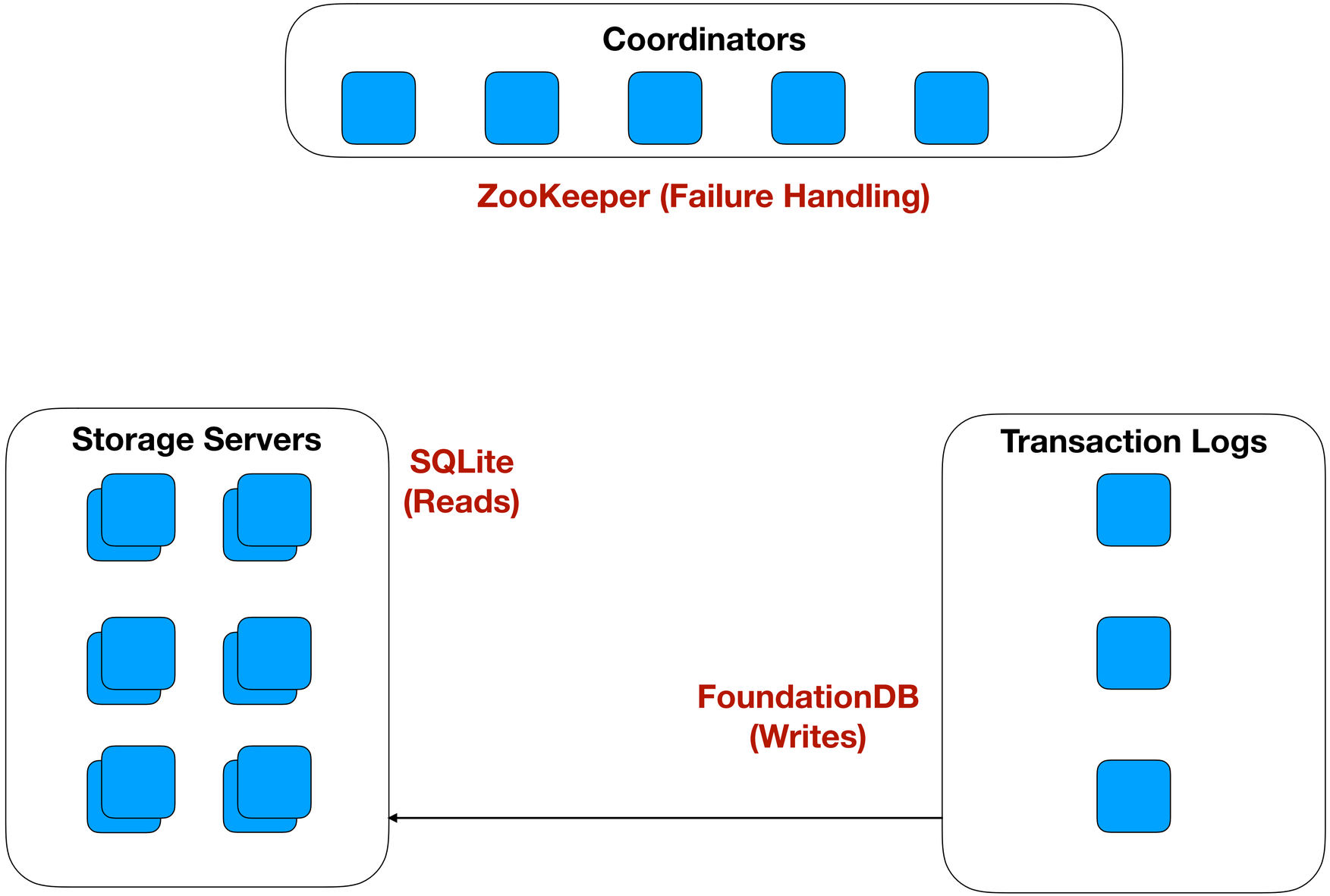

Servers
A
Writer
Reader
Wait for 2 of 3 responses
B
Wait for 2 of 3 responses
C

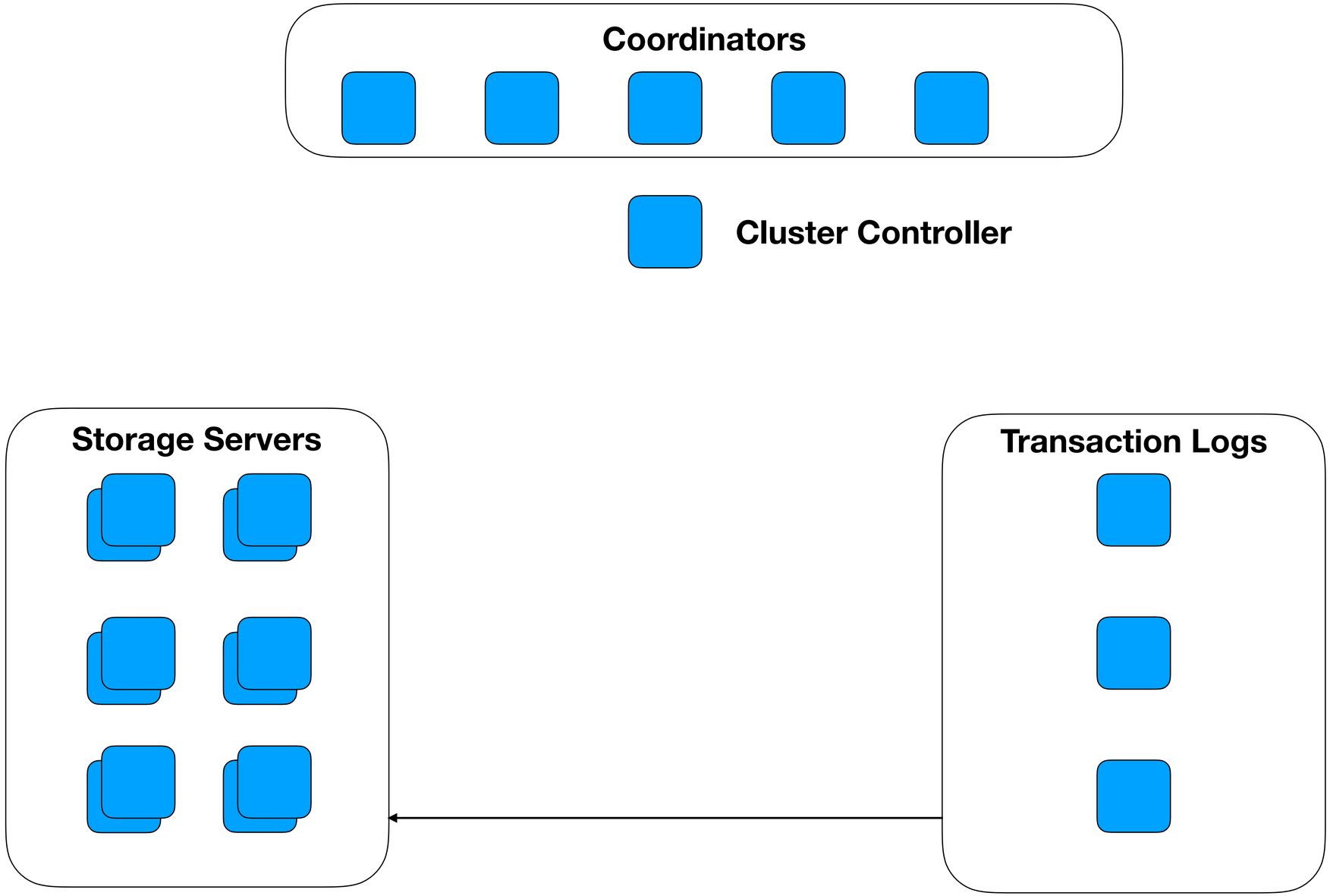

Servers
A
Writer
Reader
Wait for 3 of 3 responses
B
Wait for 1 of 3 responses
C

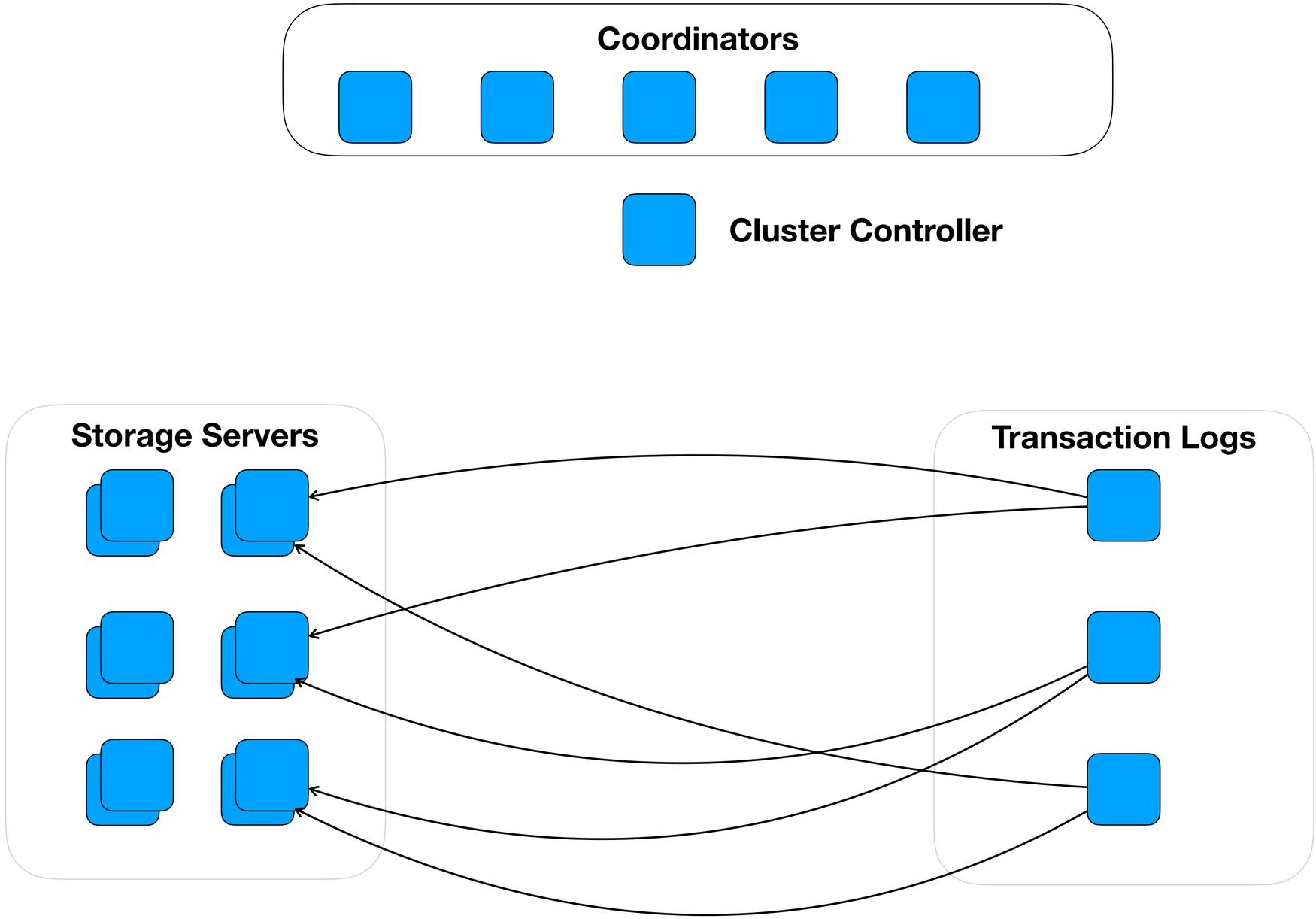

Servers
A
Writer
Reader
Wait for 3 of 3 responses
B
Wait for 1 of 3 responses
C


0: A,B,C
101: B,C,D
Versions
(Stored in a different database)

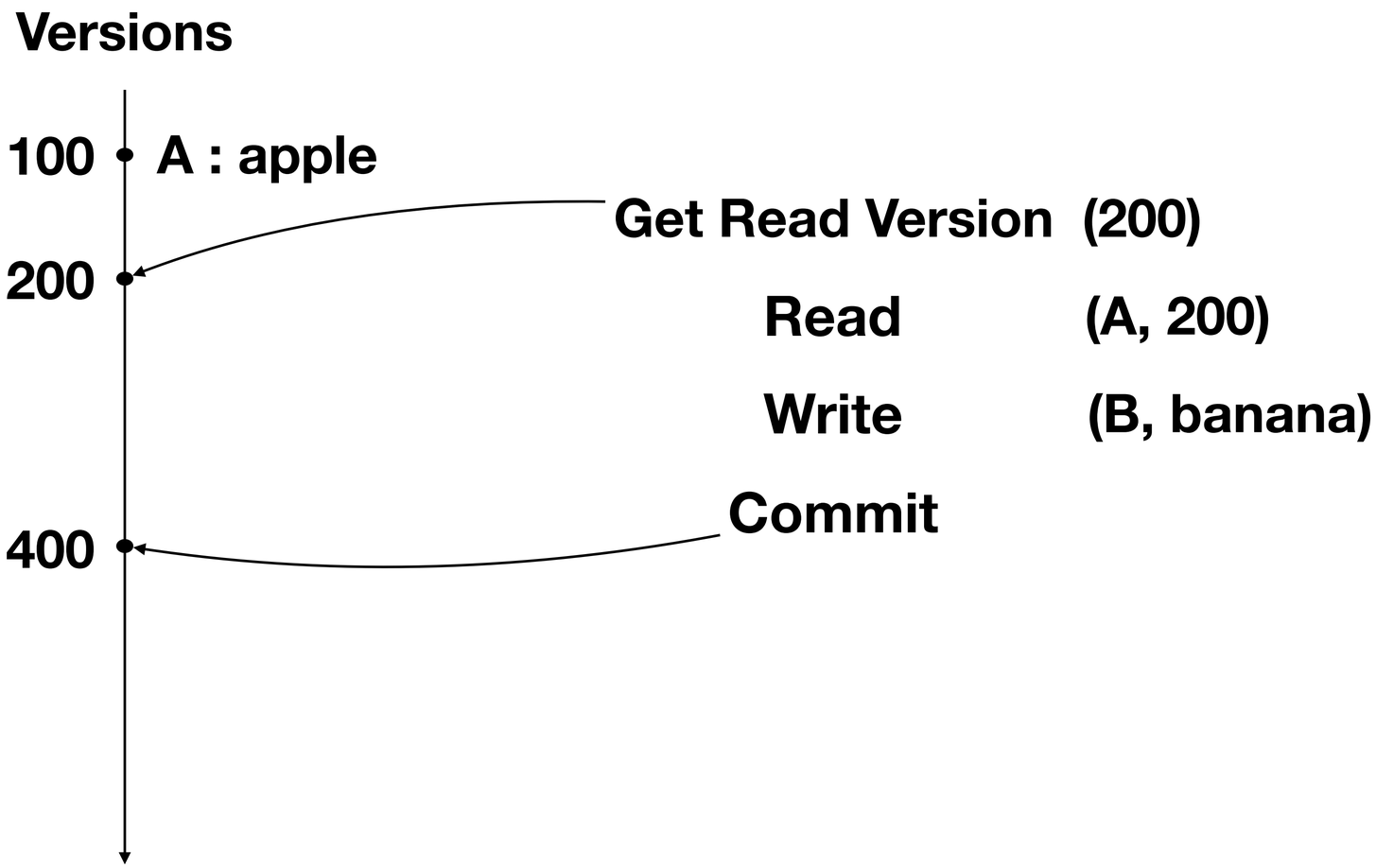

Servers
A
Writer
Reader
Wait for 3 of 3 responses
B
Wait for 1 of 3 responses
C
D


Coordinators
ZooKeeper (Failure Handling)
Storage Servers
SQLite
(Reads)
Transaction Logs
FoundationDB
(Writes)

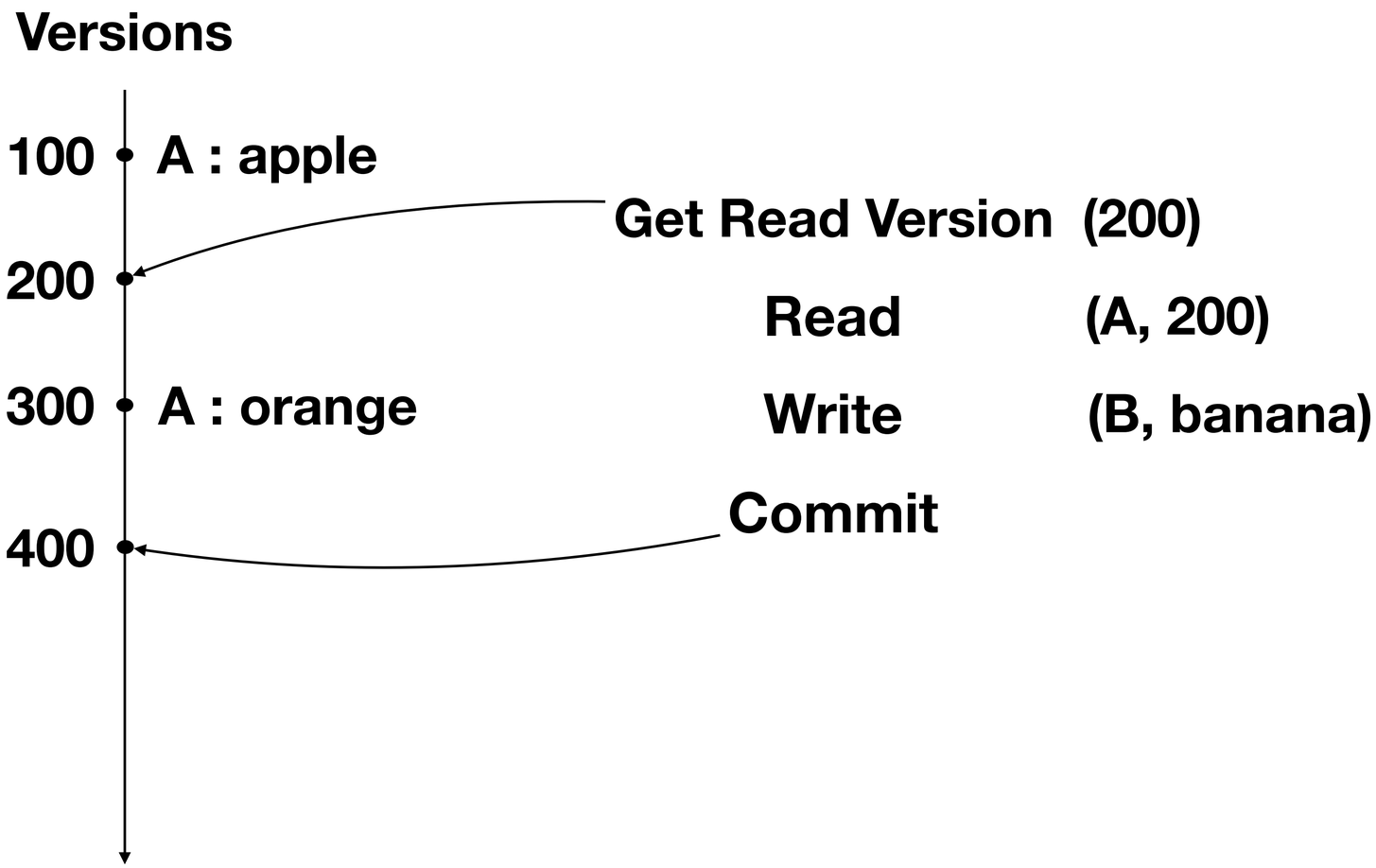

Coordinators
Cluster Controller
Storage Servers
Transaction Logs

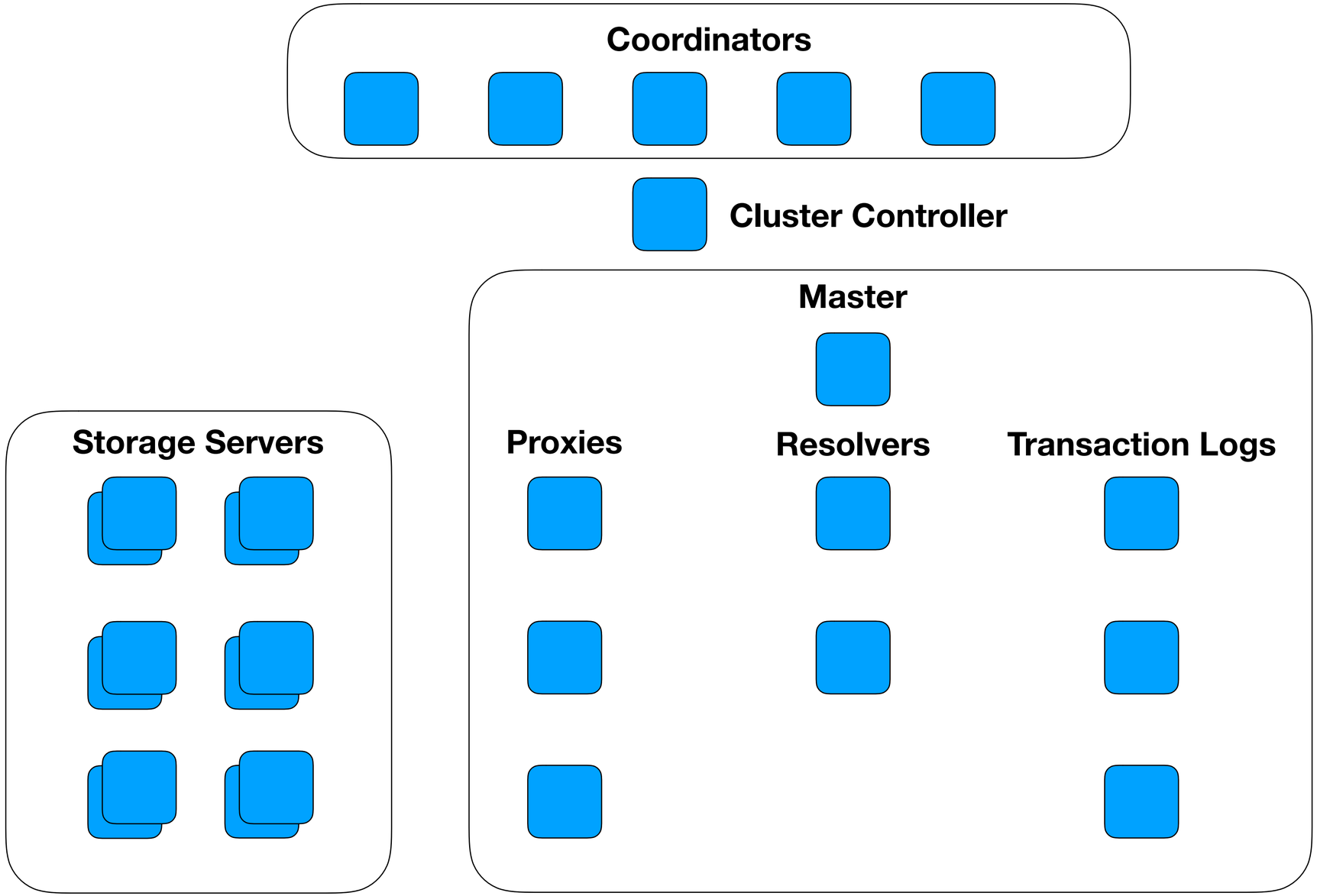

Coordinators
Cluster Controller
Storage Servers
Transaction Logs

# FoundationDB API



# FoundationDB API




Coordinators
Cluster Controller
Master
Storage Servers
Proxies
Resolvers
Transaction Logs

Get Read Version

Reads

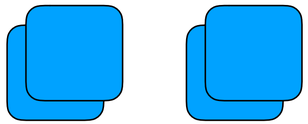
Writes

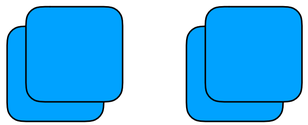
Commit

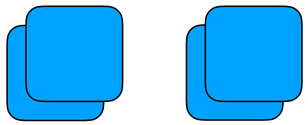
Cluster Controller

Master

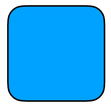
Storage Servers





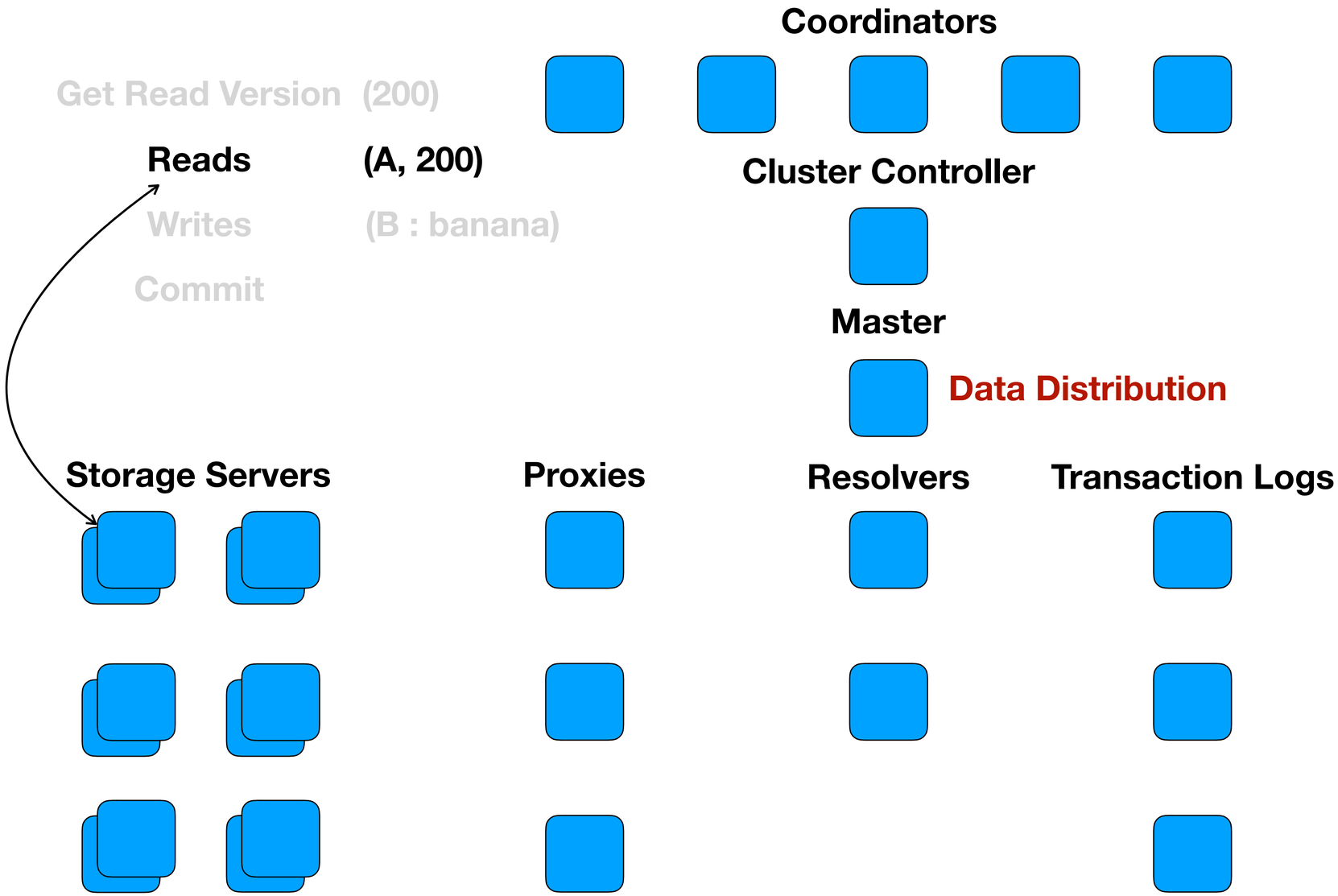


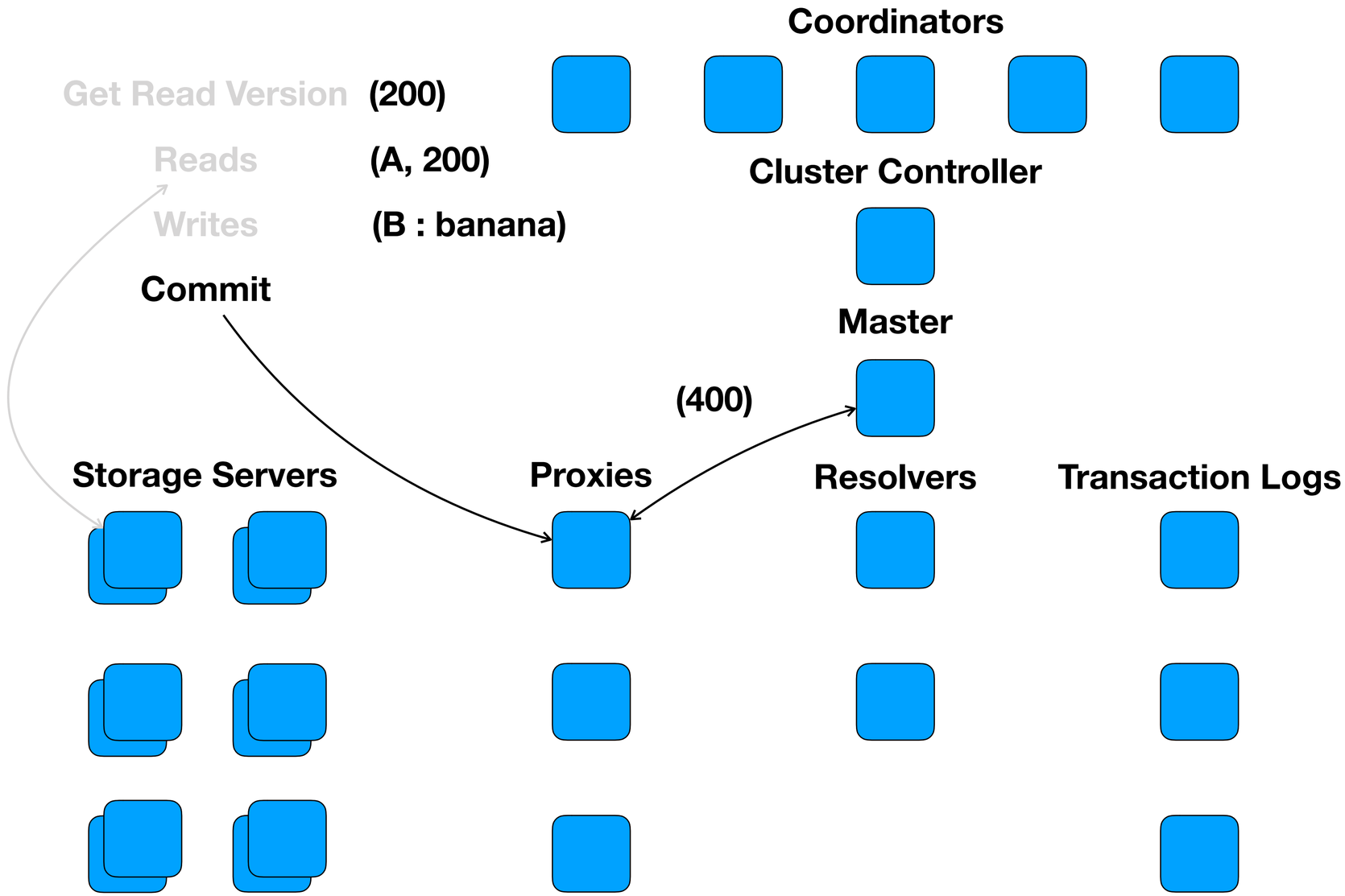
Proxies

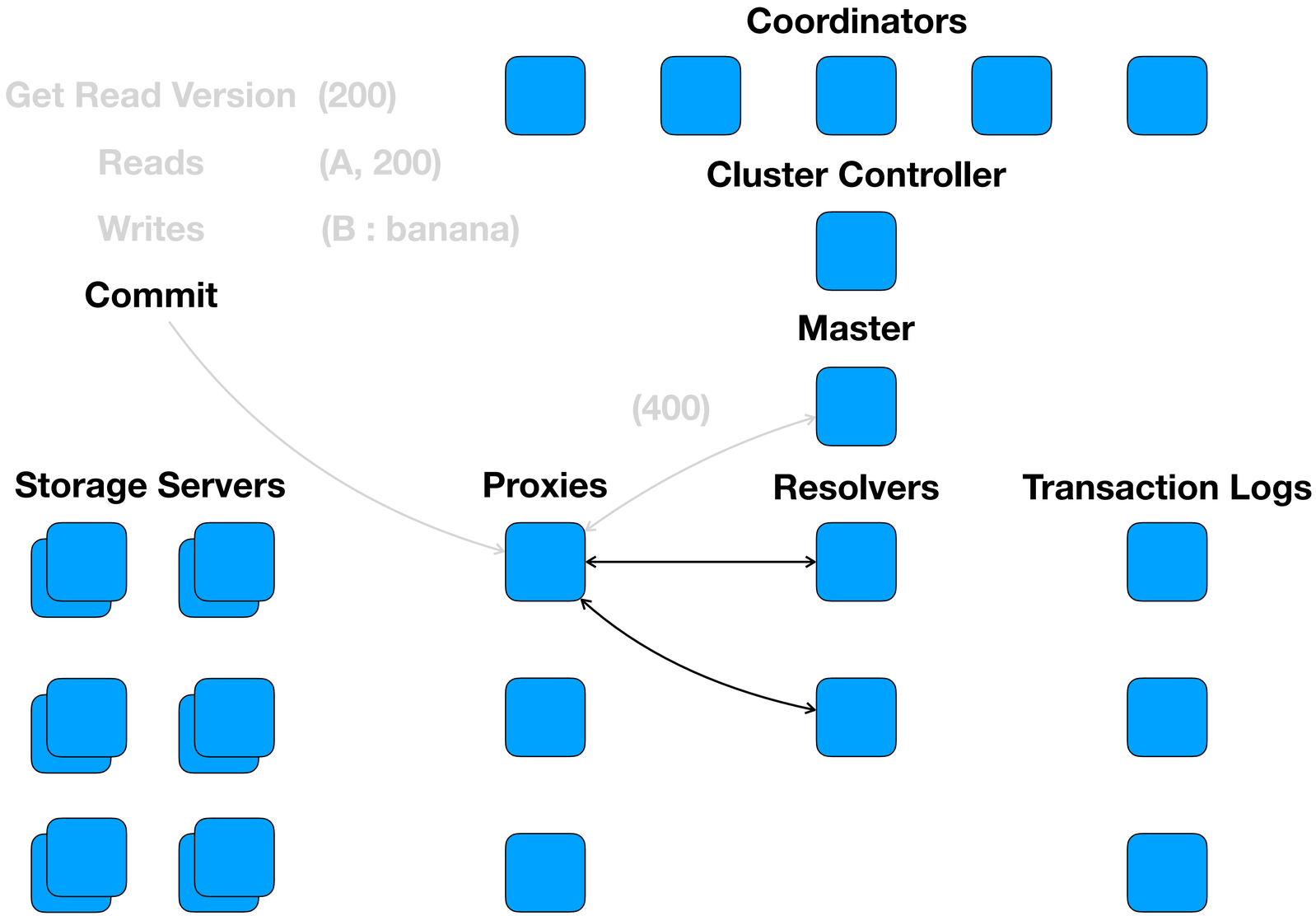


Resolvers

Transaction Logs


Coordinators
Get Read Version (200)
Reads
(A, 200)
Cluster Controller
Writes
(B : banana)
Commit
Master
Data Distribution
Storage Servers
Proxies
Resolvers
Transaction Logs


Coordinators
Get Read Version
(200)
Reads
(A, 200)
Cluster Controller
Writes
(B : banana)
Commit
Master
(400)
Storage Servers
Proxies
Resolvers
Transaction Logs

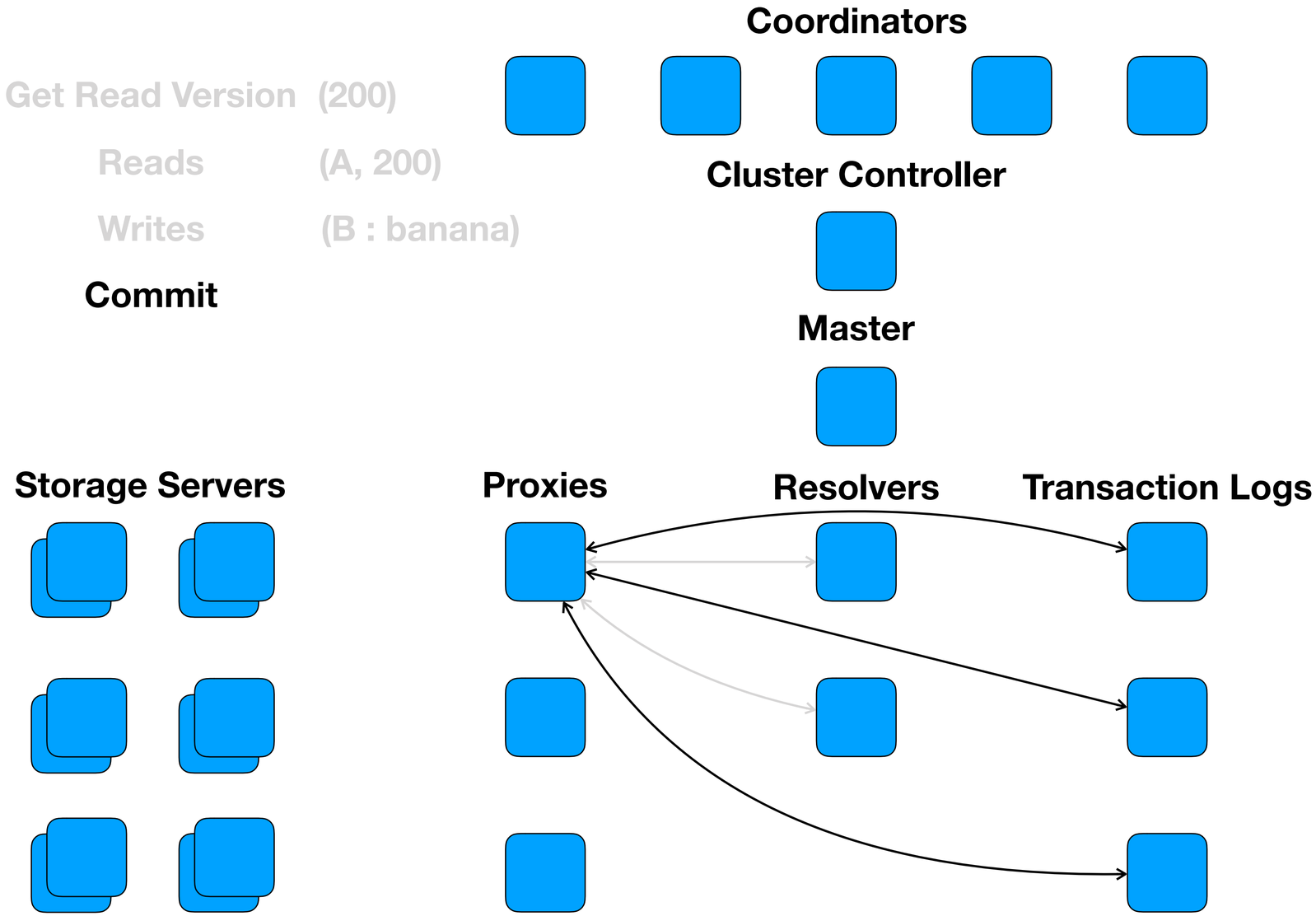

Coordinators
Get Read Version (200)
Reads (A, 200)
Writes (B : banana)
Cluster Controller
Commit
Master
(400)
Storage Servers
Proxies
Resolvers
Transaction Logs

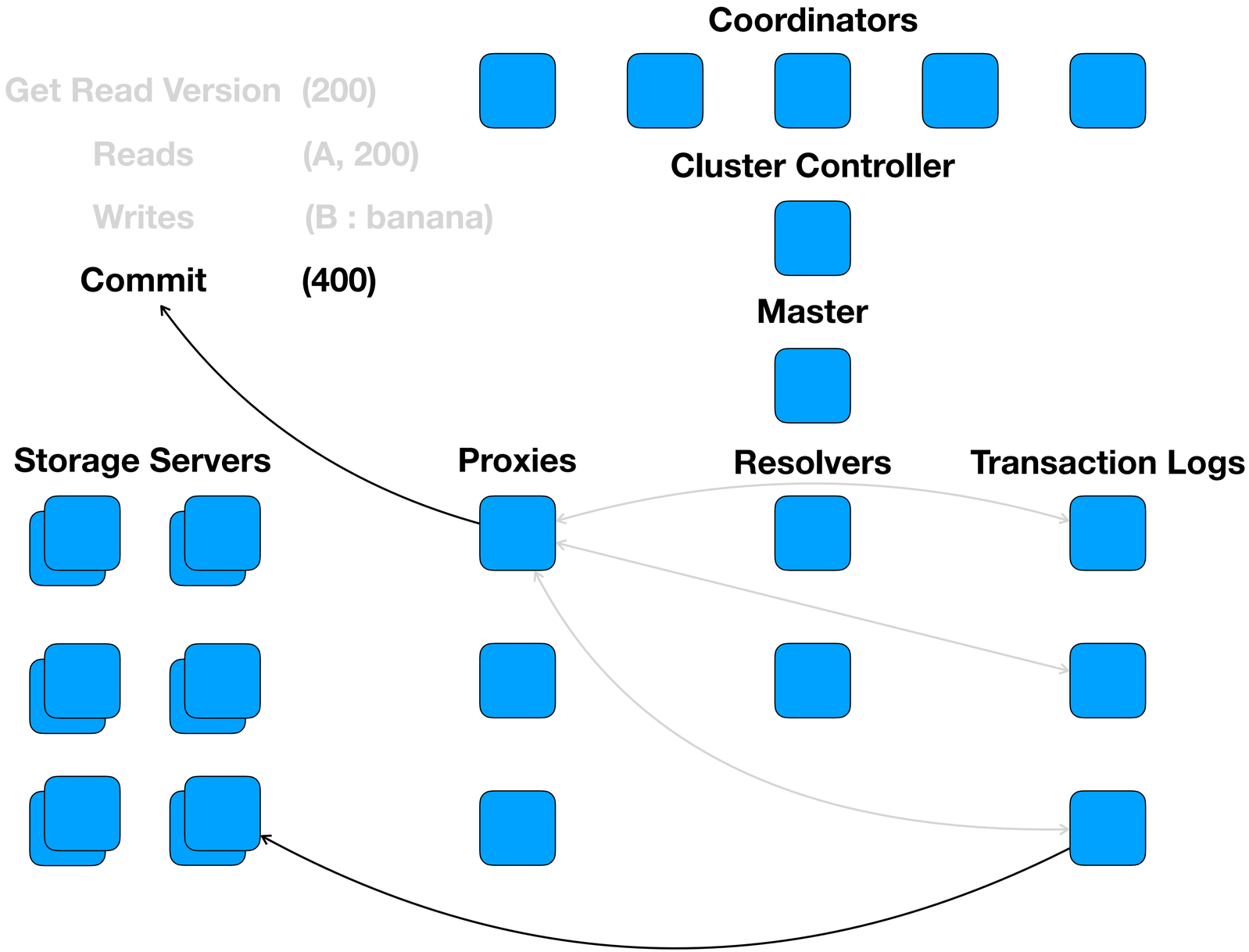
Get Read Version (200)

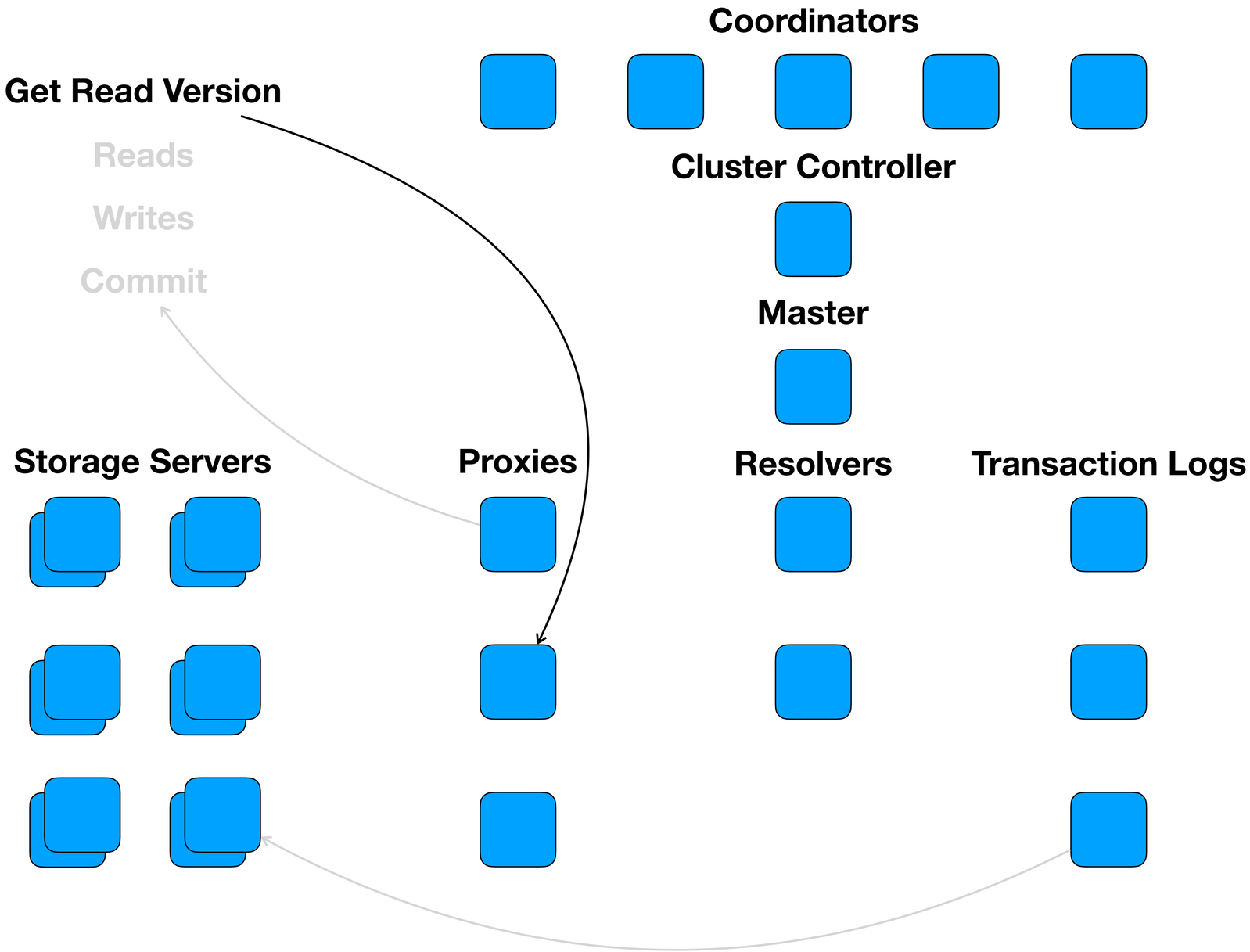
Reads (A, 200)

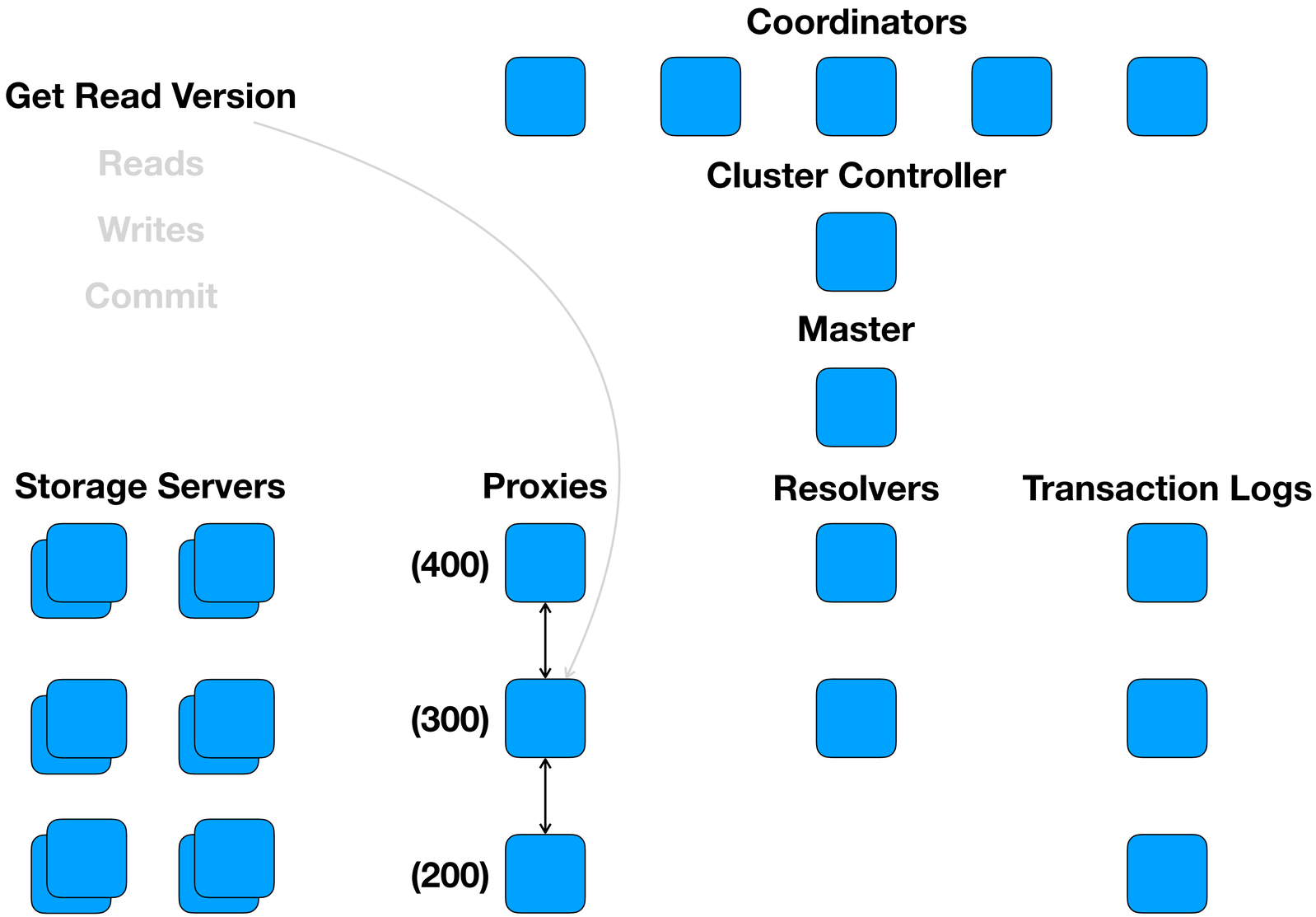
Writes (B : banana)

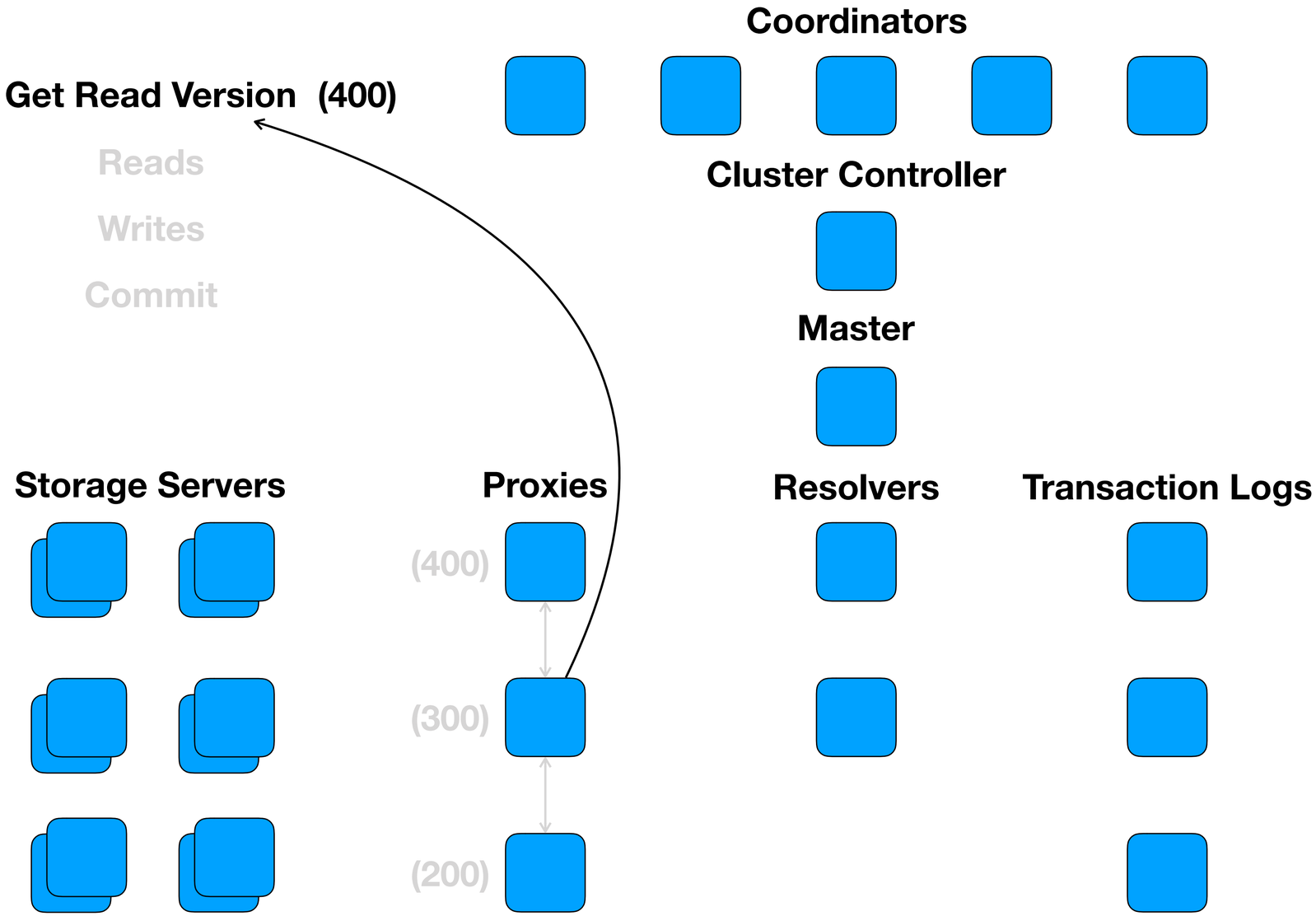
Commit




Coordinators
Get Read Version (200)
Reads (A, 200)
Writes (B : banana)
Commit (400)
Cluster Controller
Master
Storage Servers
Proxies
Resolvers
Transaction Logs


Coordinators
Get Read Version
Reads
Writes
Commit
Cluster Controller
Master
Storage Servers
Proxies
Resolvers
Transaction Logs


Coordinators
Get Read Version
Reads
Writes
Commit
Cluster Controller
Master
Storage Servers
Proxies
Resolvers
Transaction Logs
(400)
(300)
(200)

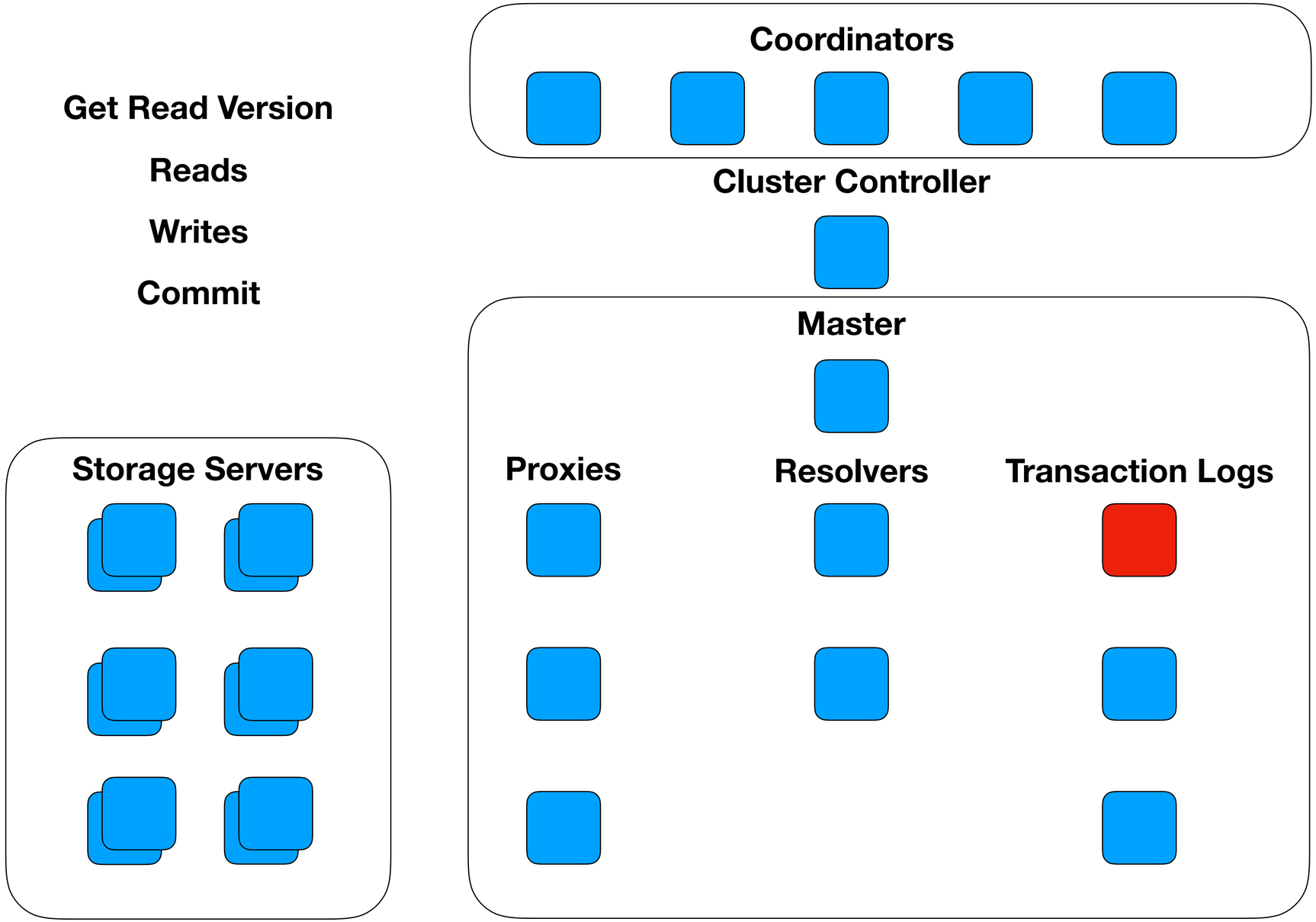

Coordinators
Get Read Version (400)
Reads
Writes
Commit
Cluster Controller
Master
Storage Servers
Proxies
Resolvers
Transaction Logs
(400)
(300)
(200)

Get Read Version

Reads

Writes

Commit

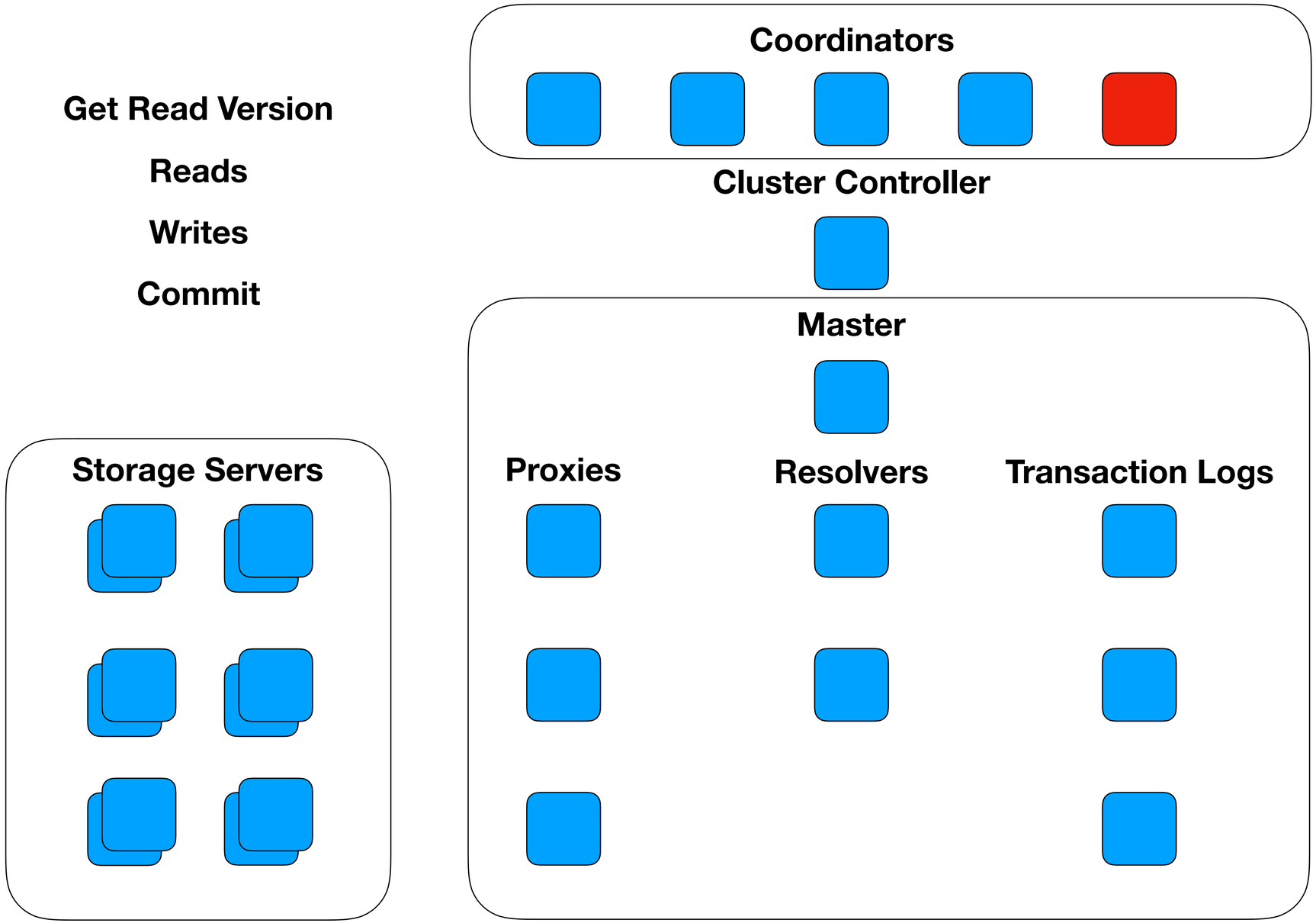


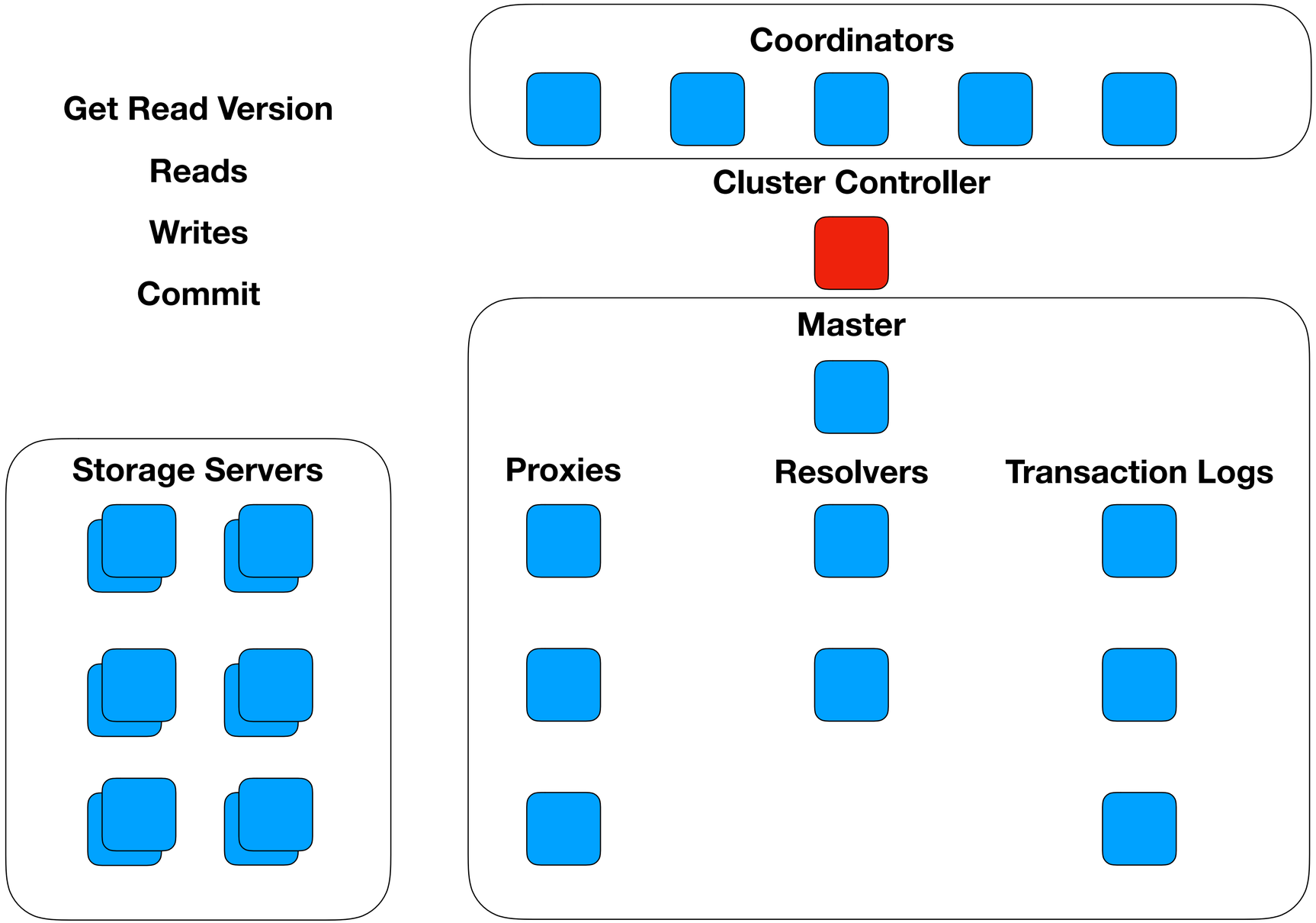
Get Read Version

Reads

Writes

Commit



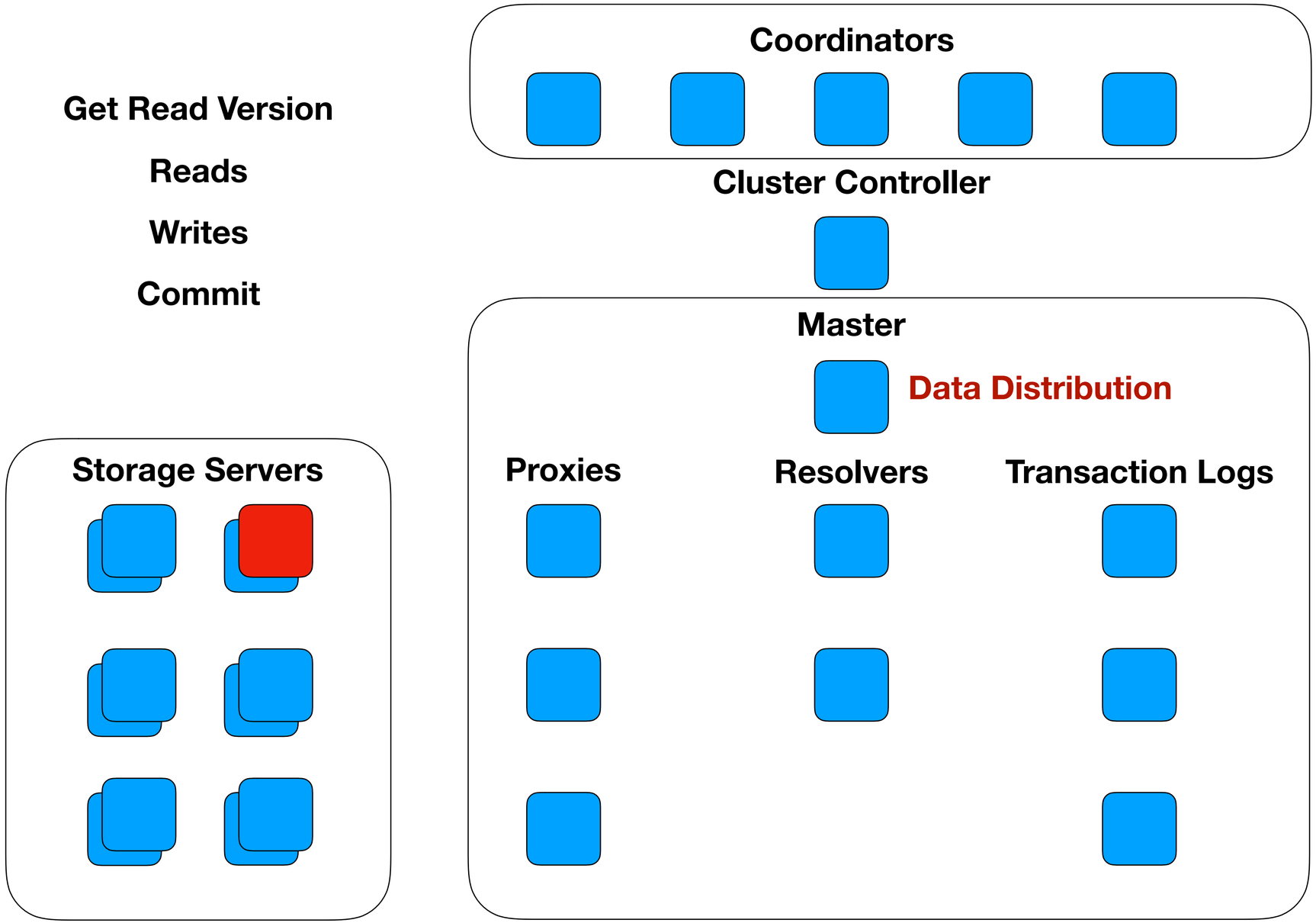

Get Read Version
Reads
Writes
Commit
Coordinators
Cluster Controller
Master
Proxies
Resolvers
Transaction Logs
Storage Servers

Get Read Version

Reads

Writes

Commit

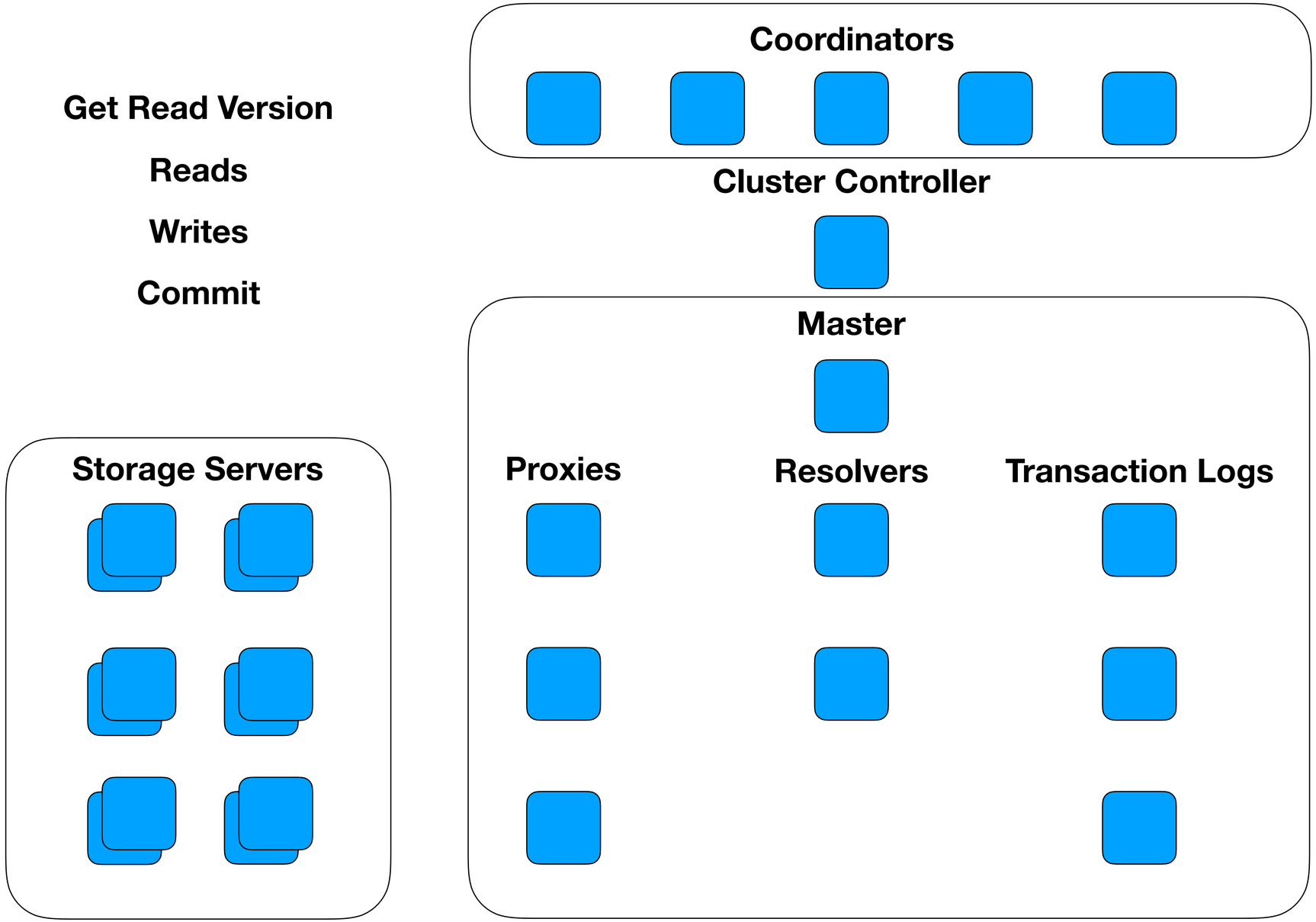


Get Read Version

Reads

Writes

Commit



# FoundationDB Performance

- Organizes many instances of a single process database into a single database
- Read and write throughput 90% of the aggregate individual read and write throughput
- Single hop read latencies
- Four hop write latencies

# Does it work?

- Run the entire distributed database in a single process
- Inject random failures
- Check correctness
- Reproduce errors

# FoundationDB

https://www.foundationdb.org